Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 2465

Polyuniverse: generation of a large-scale polymer library using rule-based polymerization reactions for polymer informatics†

Tianle Yue, Jianxin He and Ying Li *

Recent advancements in machine learning have revolutionized polymer research, leading to the swift integration of diverse computational techniques for *de novo* molecular design. A crucial aspect of these processes is to expand the number of candidate polymer structures, as the currently known real polymer structures are very limited. In contrast, small molecule databases are vast, offering extensive opportunities for the design of new molecules, such as drug discovery. In this study, we collected extensive small molecule compounds from GDB-17, GDB-13, and PubChem and selected polymerization reaction pathways for eight types of polymers, including polyimide, polyolefin, polyester, polyamide, polyurethane, epoxy, polybenzimidazole (PBI), and vitrimer. These small molecule datasets and polymerization reactions enabled us to generate hundreds of quadrillions of hypothetical polymer structures. For each of the eight polymers, along with one promising copolymer, poly(imide-imine), we randomly generated over one million hypothetical structures, except for PBI, for which we created 10 000 structures. Chemical space visualization using t-distributed stochastic neighbor embedding and synthetic accessibility scores were employed to assess the feasibility of synthesizing these new polymers. Customized feedforward neural network models predicted thermal, mechanical, and gas permeation properties for both real and hypothetical polymers. The results show that many hypothetical polymers, especially polyimides, exhibit significant potential, often surpassing real polymers in performance, particularly for high-temperature applications and gas separation. Our findings highlight the immense potential of large-scale hypothetical polymer libraries for materials discovery and design. These libraries not only aid in identifying promising polymer materials through high-throughput screening but also provide valuable datasets for training advanced machine learning models, such as large language models. This research also demonstrates the power of data-driven approaches in polymer science, paving the way for the development of next-generation polymeric materials with superior properties for diverse industrial applications.

Received 28th June 2024
Accepted 8th October 2024

DOI: 10.1039/d4dd00196f

rsc.li/digitaldiscovery

1. Introduction

Polymeric materials are ubiquitous in our daily lives, found in everything from common synthetic plastics such as polystyrene to natural biopolymers such as DNA and proteins. Their exceptional chemical, physical, biological, and mechanical properties enable a wide range of applications in the biomedical, chemical, and materials science fields.[1–5] A polymer typically consists of long chains of covalently bonded organic molecules, known as repeating units. The chemical and molecular structures of these repeating units dictate the properties of these polymeric materials.

The advancement of materials design has undergone three distinct stages. The first stage involved traditional experimentally driven and trial-and-error methods, relying heavily on experience, intuition, and conceptual insights (domain knowledge). However, this approach has inherent limitations. It provides access to only certain macroscopic properties, with many others being difficult to measure. Additionally, this method often relies on serendipitous discoveries, lacks generalizability, and is extremely time-consuming, labor-intensive, and costly. In the second stage of materials design, advances in computational technologies have led to the dominance of modeling and simulation in the field. Computational methods, such as density functional theory (DFT)[6,7] and molecular dynamics (MD),[8,9] have enabled rapid materials design through high-throughput virtual screening. These methods are particularly effective for predicting material properties when no analytical formula exists. However, computer simulations still

Department of Mechanical Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA. E-mail: yli2562@wisc.edu

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00196f

face several challenges, including high computational cost in terms of time and resources.

With the expansion of materials databases and the advancement of data science and artificial intelligence (AI) techniques, we are entering a new era often referred to as the "fourth paradigm of science"[10] or the "fourth industrial revolution."[11] This progress has ushered materials design into its third stage. Beyond experimental methods, theoretical approaches, and computer simulations, data-driven materials design has emerged as the "fourth pillar" of scientific research. Numerous breakthroughs and research efforts are now flourishing in the *de novo* design of organic molecules and polymers using data-driven methods.[12–16] Successful polymer informatics efforts have encompassed a variety of property predictions, including polymers' glass transition temperatures,[17–30] electronic bandgap,[17,31] dielectric constant,[32] and refractive index.[33] Rapidly predicting these properties enables researchers to identify optimal polymer structures with exceptional performance or those that meet specific requirements from a vast array of polymer candidates, thus facilitating the development of high-performance polymers.

However, when researchers aim to develop high-performance polymer materials using a *de novo* design strategy, rapid predictions of polymer properties through machine learning (ML) and polymer informatics are not the only requirements. A large number of candidate polymer structures are also needed for discovery and exploration. Unfortunately, the number of polymer structures in the real world is quite limited. As shown in Fig. 1, the PolyInfo dataset[34] currently includes about 18 000 experimentally synthesized polymer structures, with approximately 13 000 of these being homopolymers. In stark contrast, there is a vast number of real and hypothetical small molecule compounds. PubChem,[35] for instance, contains around 116 million real small compounds that can be purchased. Additionally, hypothetical small molecule compounds are abundant, with databases such as GDB-13 (ref. 36) and GDB-17 (ref. 37) containing nearly 977 million and 166 billion compounds, respectively. To expand the open source data for polymer informatics, Ma and Luo trained a generative model, based on the real polymer structures from PolyInfo, to generate ~1 million hypothetical polymers, namely PI1M.[38] The PI1M database spans a similar chemical space as PolyInfo but significantly populates regions where PolyInfo data are sparse.

In addition to generative models, various polymerization reactions can serve as bridges between polymer structures and small molecule compounds. Through this approach, a large number of hypothetical polymer structures with well-defined synthetic pathways can be generated based on these small molecule compounds. Simultaneously, we can examine the synthetic routes for generating these hypothetical structures, as rule-based polymerization reactions have also been validated in previous studies.

Using this strategy, Tao *et al.* generated 8 million hypothetical polyimides and discovered many polyimides with a multitude of outstanding thermal and mechanical properties.[39] By sourcing available diamine and dianhydride monomers from the PubChem database, they generated hypothetical polyimides following a predefined polycondensation reaction. To efficiently screen these compounds, they employed a ML method for high-throughput screening and evaluation. Ultimately, they identified several multifunctional polyimides that outperformed existing real polyimides and validated their properties through all-atom molecular dynamics simulations. Furthermore, these promising multifunctional polyimides were successfully synthesized based on the proposed synthetic routes, and their performance was further validated through experimental testing. Wang *et al.* generated 110 types of polyimide-derived polymer structures by combining 21 different diamine and dianhydride compounds, resulting in a wide range of electrical and thermal properties.[40] They selected 12 representative polymers, which were also successfully synthesized using the proposed synthetic routes, all derived from commercial precursors to facilitate large-scale production, and systematically investigated their structures and performance. By

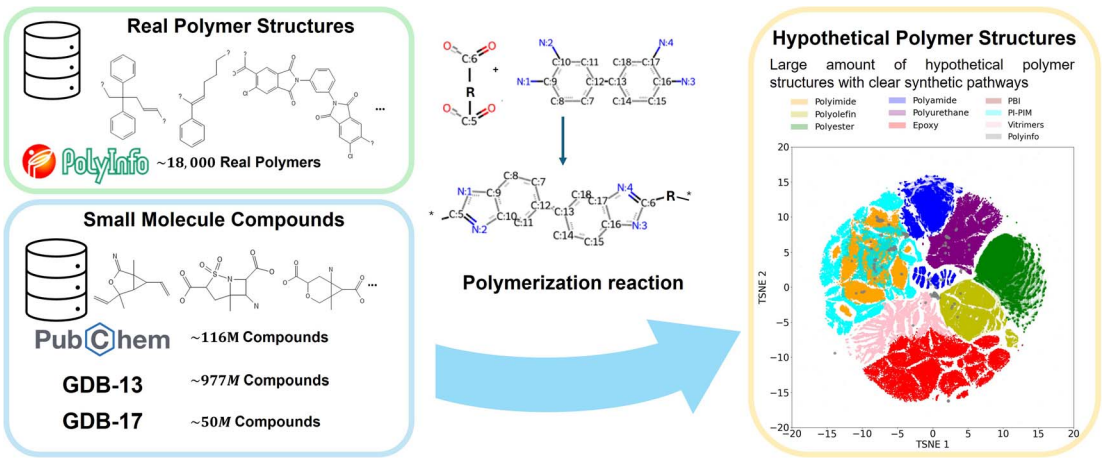

Fig. 1 Comparison of real polymer and small molecule compound datasets and the role of polymerization reactions in generating a large number of hypothetical polymer structures.

analyzing the experimental results alongside computational simulations, they quantitatively determined the impact of each structural unit on the electrical and thermal properties of the resulting polymers. This analysis revealed the key factors influencing capacitive performance at elevated temperatures for these polymers.

In addition to polyimides, Kim *et al.* developed a generative model for synthetically accessible polymer repeating units using a rule-based polymerization reaction algorithm.[41] With this system, they created a database called the Open Macromolecular Genome (OMG), which contains highly synthesizable virtual polymers. The OMG serves as an important resource for data-driven polymer research, but there is room for improvement in the definition of rule sets. From the perspective of synthetic organic chemistry, the reactivity of a substrate is influenced by the steric and electronic effects of substituents at the reaction center. Additionally, as highlighted in their work, the selectivity of the reaction is affected by coexisting functional groups in the reactant molecule. Therefore, it is necessary to develop reaction rules that account for these factors. Ohno *et al.* developed a virtual library generator for polymers that incorporates a comprehensive rule set for practically applied polymerization reactions using a Python open-source library called Small Molecules into Polymers (SMiPoly).[42] This generator implements 22 reaction rules, which include six chain polymerization reactions and 16 step-growth polymerization reactions. Overall, the system enables the synthesis of seven different types of polymers. Additionally, Ferrari *et al.* used large language models and fine-tuned the polymerization models for both forward and backward prediction tasks, addressing both homo-polymers and co-polymers consisting of up to two monomers. Their model predicts reactants, as well as reagents, solvents, and catalysts for each step of the retro-synthesis.[43]

However, previous studies based on polymerization reactions have either focused on only one specific type of polymer or on developing efficient algorithms for generating hypothetical polymers, often neglecting the analysis and property prediction of large-scale hypothetical polymer structures generated from various types of polymerization reactions. Therefore, in this study, we aim to generate a wide range of hypothetical polymer structures using polymerization reactions, targeting multiple popular or promising classes of polymers, and subsequently analyze and predict their properties through machine learning techniques.

In this study, we selected eight popular and promising types of polymers—polyimide, polyolefin, polyester, polyamide, polyurethane, epoxy, polybenzimidazole (PBI), and vitrimers—along with one promising copolymer, poly(imide-imine) (PI-PIM). Hundreds of quadrillions of hypothetical polymer structures can be generated based on small molecule compounds from the GDB-17, GDB-13, and PubChem datasets and well-defined polymerization reactions. For each type of polymer, we randomly generated 1 million hypothetical structures, except for PBI, for which only 10 000 hypothetical structures were generated. The chemical space location of all generated polymers was obtained, and the synthetic accessibility (SA) score provides an estimation of their synthesis difficulty. Then, ML methods are employed to predict various thermal and mechanical properties, as well as several types of gas permeabilities. The distribution of the prediction results reveals the distinct characteristics of different types of polymers. To demonstrate the potential of the large number of hypothetical polymer structures generated, we also identified the best real polymer provided by PolyInfo and compared it to hypothetical polymer structures that outperformed it. These results showcase that many hypothetical polymers, especially polyimides, exhibit significant potential, often surpassing real polymers in performance, particularly for high-temperature applications and gas separation.

2. Results & discussion

2.1 Polymer class

The correlation between the molecular structure and properties is pivotal for advancing polymer science and engineering. This research initiative has established a comprehensive database of polymer structures to support innovations in their application and development. The database encompasses a variety of polymer types, each selected for its unique properties that are essential for broad industrial applications.

For example, polyimides are recognized for their thermal stability, derived from aromatic backbones and imide functionalities, making them suitable for high-temperature environments. Similarly, polyurethanes, with their segmented block copolymer structure, are crucial for automotive and construction applications. Additionally, PI-PIMs exhibit rehealability and recyclability enabled by dynamic imine bonds, while retaining the excellent mechanical and thermal properties of polyimide.[44] These examples highlight how specific microstructural characteristics critically determine the functionalities of these polymers.

Here, a large-scale library of polymer structures was generated by applying specific polymerization reactions. Guided by the fundamental principles of polymerization,[45,46] condensation reactions were used to generate polyimides, polyamides, polyurethanes, polyesters, PBIs, and PI-PIMs *via* step-growth mechanisms that link monomers and facilitate the removal of small molecules. Ring-opening reactions were employed to produce epoxy and vitrimers, transforming cyclic monomers into network structures. Additionally, both single and dual monomer addition polymerizations were implemented for polyolefins, capturing a spectrum from simple linear polymers to complex copolymers. Monomers were selected based on the necessary functional groups for these polymerizations, ensuring that the dataset accurately reflects a diverse array of polymer structures and aligns with specific synthesis pathways, as depicted in Fig. 2 and Table 1.

2.2 Small molecule compound datasets

The small molecule compounds used to generate specific types of hypothetical polymers based on the polymerization reactions were selected from the GDB-17, GDB-13, and PubChem databases according to the functional groups required. GDB-13 and

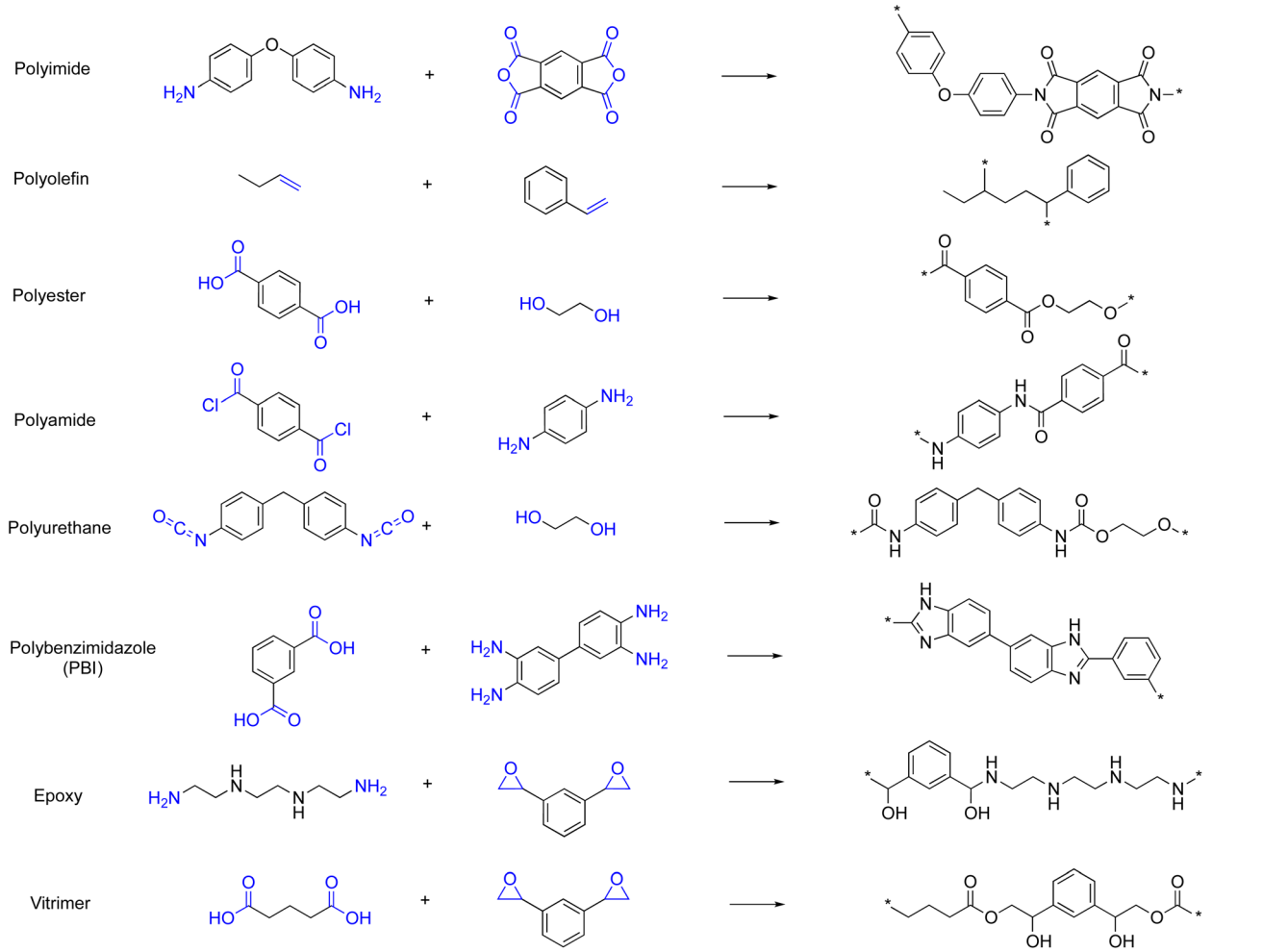

Fig. 2 Examples of generated polymers from small molecule compounds for each polymer class—polyimide, polyolefin, polyester, polyamide, polyurethane, epoxy, polybenzimidazole, and vitrimers—along with their polymerization reactions. For vitrimers, only the reaction between epoxides and carboxylic acids is used because these two functional groups are common and abundant.

Table 1 Selected polymer types and corresponding small molecule compounds used for synthesis

Polymer class	Monomer class
Polyimide	Polycarboxilic acid anhydride and polyamine
Polyolefin	Vinylidene and cyclic olefin
Polyester	Lactone, hydroxy carboxylic acid, polyol and thiol, carbon monoxide, polycarboxylic acid and acid halide, and epoxide
Polyamide	Lactam, amino acid, polycarboxylic acid and acid halide, and polyamine
Polyurethane	Polyisocyanate, polyol and thiol
Epoxy	Epoxide and polyamine
PBI	Polycarboxylic acid and acid halide and 3,3′,4,4′-tetraaminodiphenyl
Vitrimers	Epoxide, polycarboxylic acid and acid halide

GDB-17 are extensive datasets of hypothetical small molecules. GDB-13 includes molecules containing up to 13 atoms of carbon, nitrogen, oxygen, sulfur, and chlorine, following rules for chemical stability and synthetic feasibility, comprising 977 468 314 structures.[36] GDB-17 extends this enumeration to molecules with up to 17 atoms of carbon, nitrogen, oxygen, sulfur, and halogens, resulting in a total of 166.4 billion molecules, with only 50 million structures publicly available.[37] PubChem is an open chemistry database maintained by the National Institutes of Health (NIH). PubChem contains a vast array of chemical data, including small molecules, nucleotides, carbohydrates, lipids, peptides, and chemically modified macromolecules. It provides comprehensive information on chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, and toxicity data.[35]

GDB-17 and GDB-13 were chosen because they offer extensive coverage of chemical space, and PubChem was selected because it contains easily accessible real small compounds. Besides these three chosen datasets, there are many other small molecule datasets available for researchers, such as ChEMBL,[47] ZINC,[48] ChemSpider,[49] and DrugBank.[50] These datasets can also be used to generate hypothetical polymer structures. The selected small molecules include amino acids, cyclic olefins, epoxides, hydroxy carboxylic acids, lactams, lactones, polycarboxylic acids and acid halides, polyamines, polycarboxylic acid anhydrides, polyisocyanates, polyols and thiols, and vinylidenes. Fig. 3 illustrates the quantities of these small molecule compounds within the three small molecule datasets, respectively (see ESI Table S1 for detailed counts and ESI Tables S2–S4† for information about more functional groups).

From Fig. 3, it is evident that the GDB-13 database contains a significantly higher quantity of cyclic olefins, polyamines, and vinylidene monomers compared to other compounds. Overall, GDB-13 appears to have the highest overall quantity of small molecules, which is closely related to the fact that the GDB-13 dataset contains significantly more small molecules than the other two datasets. The GDB-17 dataset theoretically should include far more small molecules than GDB-13, but currently, only 50 million have been made publicly available. This also makes the distribution of the GDB-17 dataset appear somewhat more balanced compared to GDB-13. The GDB-13 and GDB-17 datasets both have relatively low quantities of polycarboxylic acids and acid halides. Furthermore, it is also important to note that there are some small molecules missing from the GDB-17 and GDB-13 datasets. GDB-13 does not include any polyisocyanates. Additionally, GDB-17 lacks not only this type of small molecule but also polycarboxylic acid anhydrides.

The PubChem database, however, shows a more balanced distribution across different compounds. The balanced distribution in the PubChem dataset is due to its source, as it collects a wide variety of small molecules that are both real and purchasable. This balanced distribution is especially important given the absence of certain types of small molecules in the GDB-13 and GDB-17 datasets. However, we can observe that, similar to the previously mentioned GDB-13 and GDB-17 datasets, the PubChem dataset also has relatively low quantities of polycarboxylic acid anhydrides and polyisocyanates.

Table 2 shows the total number of unique structures for each type of small molecule from the three datasets, representing the variety of molecules that are readily available for use. This distribution of small molecules across these databases highlights their utility in generating diverse hypothetical polymer structures for further research. They can provide an enormous number of hypothetical polymer structures. For example, polyimides, which can be generated from polycarboxylic acid anhydride and polyamine small molecule compounds, have 9253 polycarboxylic acid anhydrides and 207 640 913 polyamines are available. This means that we can generate approximately 2 trillion hypothetical polyimide structures. Similarly, polyolefins, which can be generated from vinylidene and cyclic olefin small molecule compounds, have 193 219 664 vinylidenes and 207 640 913 cyclic olefins available. This allows for the generation of around 120 quadrillion hypothetical polyolefin structures. However, for PBI, which can be generated from polycarboxylic acid and acid halide and 3,3′,4,4′-tetraaminodiphenyl, there are only 550 440 polycarboxylic acid and acid halide monomers available. As a result, the number of hypothetical PBI structures that can be generated is relatively limited. Table 3 shows the theoretical maximum number of hypothetical structures generated for each polymer class using the three small molecule datasets previously described.

These vast quantities of hypothetical polymer structures have immense potential for utilization. Researchers can use

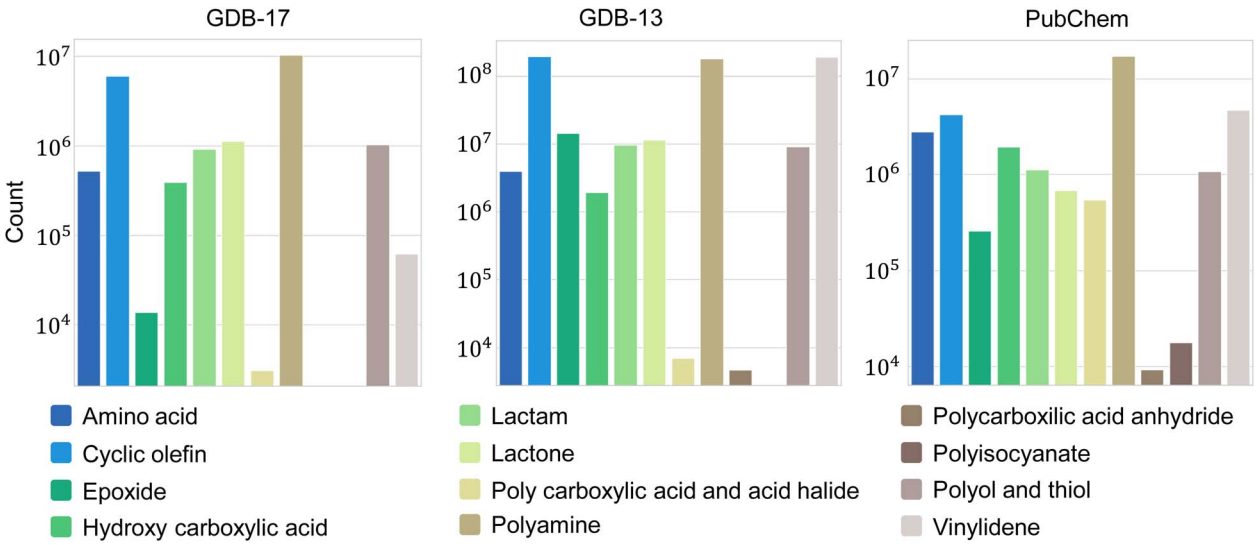

Fig. 3 Quantities of small molecule compounds within the three datasets, including amino acids, cyclic olefins, epoxides, hydroxy carboxylic acids, lactams, lactones, polycarboxylic acids and acid halides, polyamines, polycarboxylic acid anhydrides, polyisocyanates, polyols and thiols, and vinylidenes.

Table 2 Number of unique structures for each type of small molecule from the three datasets, GDB-13, GDB-17, and PubChem

Monomer class	Count	Monomer class	Count
Amino acid	7 256 230	Polycarboxylic acid and acid halide	550 440
Cyclic olefin	204 472 259	Polyamine	207 640 913
Epoxide	14 825 849	Polycarboxylic acid anhydride	9253
Hydroxy carboxylic acid	4 226 491	Polyisocyanate	17 631
Lactam	11 626 974	Polyol and thiol	14 676 768
Lactone	13 266 515	Vinylidene	193 219 664

Table 3 Theoretical maximum number of hypothetical structures generated for each polymer class using three small molecule datasets, GDB-13, GDB-17, and PubChem

Polymer class	Theoretical maximum number
Polyimide	1 921 301 367 989
Polyolefin	120 568 894 750 259 696
Polyester	18 136 241 831 465
Polyamide	166 946 749 591 594
Polyurethane	258 766 096 608
Epoxy	3 078 452 822 360 137
PBI	550 440
Vitrimers	8 160 740 323 560

high-throughput screening methods to identify promising polymer materials. Additionally, they can be employed to train generative models or large language models, as these ML models require extensive polymer structure information for training data. Furthermore, since we also have the polymerization reaction pathways and small molecule information for these hypothetical polymer structures, combining them with polymer informatics offers even more possibilities for researchers.

2.3 Generation of hypothetical polymer structures

Using the polymerization reaction pathways and small molecule datasets, we randomly selected small molecules and generated 1 million hypothetical polymer structures for each type of polymer, except for PBI, for which we generated 10 thousand hypothetical polymer structures. Fig. 4(a) illustrates the chemical space visualization of real polymers from the PolyInfo dataset along with all the hypothetical polymers for each type of polymer as well as PI-PIM. T-distributed Stochastic Neighbor Embedding (TSNE) is a technique used for embedding high-dimensional data into two-dimensional spaces.[51] TSNE is a popular nonlinear dimensionality reduction and data visualization method that preserves nonlinear similarities between data points. It works by first calculating the similarity between high-dimensional data points using a Gaussian distribution, then calculating the similarity between data points in the low-dimensional space using a t-distribution, and finally minimizing the difference between the high-dimensional and low-dimensional similarities. It is evident that the structures of each type of polymer are relatively clustered in the chemical space, with each polymer type generally occupying a specific region. Additionally, since PI-PIM is a copolymer that includes polyimide, its chemical space overlaps with that of polyimide. On the other hand, the real polymers in the PolyInfo dataset encompass many types, resulting in a much more dispersed distribution throughout the chemical space.

Furthermore, we incorporated the SA score index to assess the feasibility of synthesizing these hypothetical polymers. The SA score index is a method that characterizes the synthetic accessibility of molecules, assigning a score between 1 (easy to make) and 10 (very difficult to make). Fig. 4(b) illustrates the SA score distributions of all the hypothetical polymers for each type

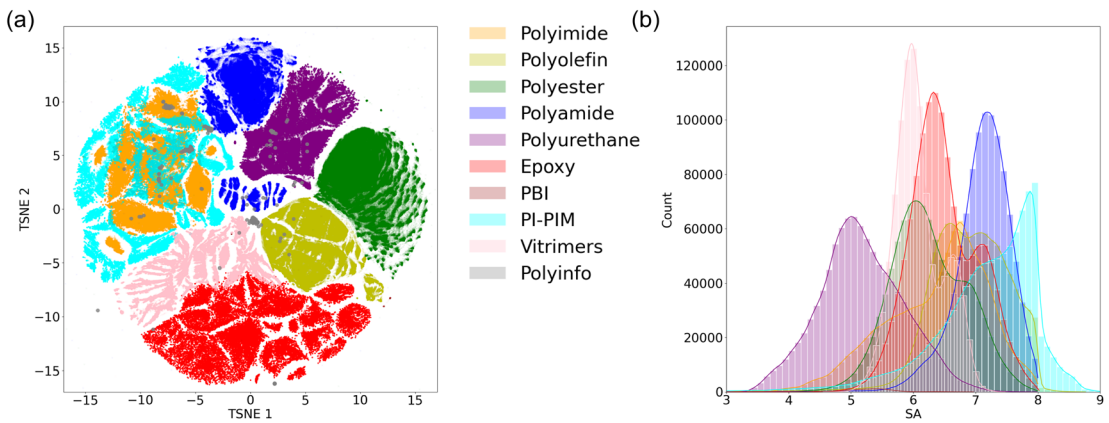

Fig. 4 (a) Chemical space visualization and (b) SA score distributions of the real polymer data set from PolyInfo and generated hypothetical polyimide, polyolefin, polyester, polyamide, polyurethane, epoxy, PBI, PI-PIM, and vitrimers.

of polymer as well as PI-PIM. It can be seen that most of the hypothetical polymer structures have SA scores ranging between 4 and 8. It is important to note that the calculation of the SA score is highly related to the complexity of the small molecules. In this study, the use of a large number of small molecule compounds from GDB-13 and GDB-17 resulted in higher SA scores for the hypothetical polymer structures. If the goal is to obtain more easily synthesizable hypothetical polymer structures, using small molecule compounds solely from PubChem would be feasible.

2.4 ML for high-throughput screening of real and hypothetical polymer datasets

We then implemented customized feedforward neural network (FNN) models, based on our previous benchmark study,[52] to screen the real polymer dataset (PolyInfo) and all generated hypothetical polymers, with a particular focus on thermal, mechanical, and gas permeation properties, based on our previous studies.[39,52–55] For thermal properties, we predicted glass transition temperature (T_g), melting temperature (T_m), and decomposition temperature (T_d). For mechanical properties, we predicted Young's modulus (E), yield strength (σ_y), and breaking strength (σ_b). For gas permeation properties, we focused on six gases: helium (He), hydrogen (H_2), oxygen (O_2), nitrogen (N_2), carbon dioxide (CO_2), and methane (CH_4).

The polymer structures were represented using polymer-simplified molecular input line entry system (p-SMILES) strings generated using RDKit.[56] In this system, SMILES strings were used to define the structures of the repeat units, and a pair of asterisks ('*') was employed to indicate the two endpoints of the repeat unit, representing the polymerization points. For predicting the three thermal properties, the Morgan Fingerprint with Frequency (MFF), which is efficient and robust in generating an interpretable molecular representation of polymers,[52,53] was employed as the input to the FNN model. The datasets for T_g, T_m, and T_d are detailed in ESI Fig. S1,† and the training results for T_g, T_m, and T_d are detailed in ESI Fig. S2.† For mechanical properties except σ_b and gas permeation properties, models from our previous work were used for predictions.[54,55] The dataset for E, σ_y, and σ_b are detailed in ESI Fig. S3,† and the training results for E, σ_y, and σ_b are detailed in ESI Fig. S4.† The T_g prediction results are validated with molecular dynamics simulation and detailed in ESI† titled "Details of molecular dynamics verification."

2.4.1 Thermal properties. Thermal properties of polymers, such as T_g, T_m, and T_d, are crucial for several reasons. These properties determine the polymer's behavior and stability under different temperature conditions, which directly impact their mechanical performance, processing, and safety. T_g is a critical property that controls the phase transition of polymers, thereby influencing their applications.[57] T_m defines the processing conditions, allowing for shaping and forming of the polymer. T_m is one of the most important thermal properties of polymers, as it significantly impacts both the thermodynamics and physical behavior of the polymer during processing, as well as the final morphology of the product. Several studies have noted that the values of T_m and T_g are generally approximately proportional to each other.[57] T_d provides information on the polymer's thermal stability and safety, ensuring it does not degrade prematurely. Once T_d is exceeded, the molecular structure of the polymer irreversibly changes, leading to a significant decline in its properties or the production of gases, liquids, or other chemical byproducts.

In general, for most polymers, T_g is lower than the T_m because T_g primarily involves the movement of polymer chain segments, while T_m corresponds to the melting of the entire structure.[58] Furthermore, T_d of polymers is typically higher than T_m, as chemical decomposition generally requires more energy than the melting of polymer chains. Understanding these thermal properties helps us in selecting appropriate polymers for various applications, optimizing manufacturing processes, and ensuring the material's performance and longevity.

Fig. 5(a)–(c) display the distribution of T_g, T_m, and T_d prediction values for real polymers from PolyInfo and for each type of generated hypothetical polymer. It can be observed that for each type of polymer, the predicted values for the three thermal properties are quite continuous, with most displaying a near-Gaussian distribution. This aligns with the distribution of polymer property values in the real world. By comparing the predicted results across different types of polymers, it is evident that the predicted value range for polyimides is higher than that for other types of polymers. A significant number of hypothetical polyimide structures are distributed in the high-temperature region (>300 °C). This observation aligns with real-world knowledge that polyimides are high-performance engineering plastics known for their excellent strength and stiffness, exceptional heat resistance, and chemical stability. Their attractive mechanical and thermal properties are widely utilized in the aerospace, automotive, and electronics industries.[59–64] Some polyimides can withstand temperatures of up to 400 °C and maintain excellent mechanical properties across a broad temperature range (−269 °C to 400 °C).[39]

Fig. 5(d) displays the structure of the real polymer with the highest combined predicted values of T_g, T_m, and T_d from the PolyInfo dataset (shown within the gray box), alongside the structure with the highest combined predicted values from all generated hypothetical polymer structures. (The T_g prediction results are validated using molecular dynamics simulation and are detailed in the ESI† section titled "Details of Molecular Dynamics Verification.") This top-performing structure comes from the 1 million hypothetical polyimide structures. The radar chart compares their predicted performance, and on the far right, the small molecule compounds used to synthesize this hypothetical polyimide structure are shown. It is evident that the predicted performance of this hypothetical polyimide structure surpasses that of the real polymer in all aspects, showcasing the potential of these hypothetical polymer structures for high-temperature applications. Additionally, an explainable machine learning technique, SHapley Additive exPlanations (SHAP) analysis,[65] was further employed to evaluate the impact of substructures on the T_g of the hypothetical polyimide (ESI Fig. S5†). The SHAP analysis revealed that the high T_g value of the promising structure is primarily due to the

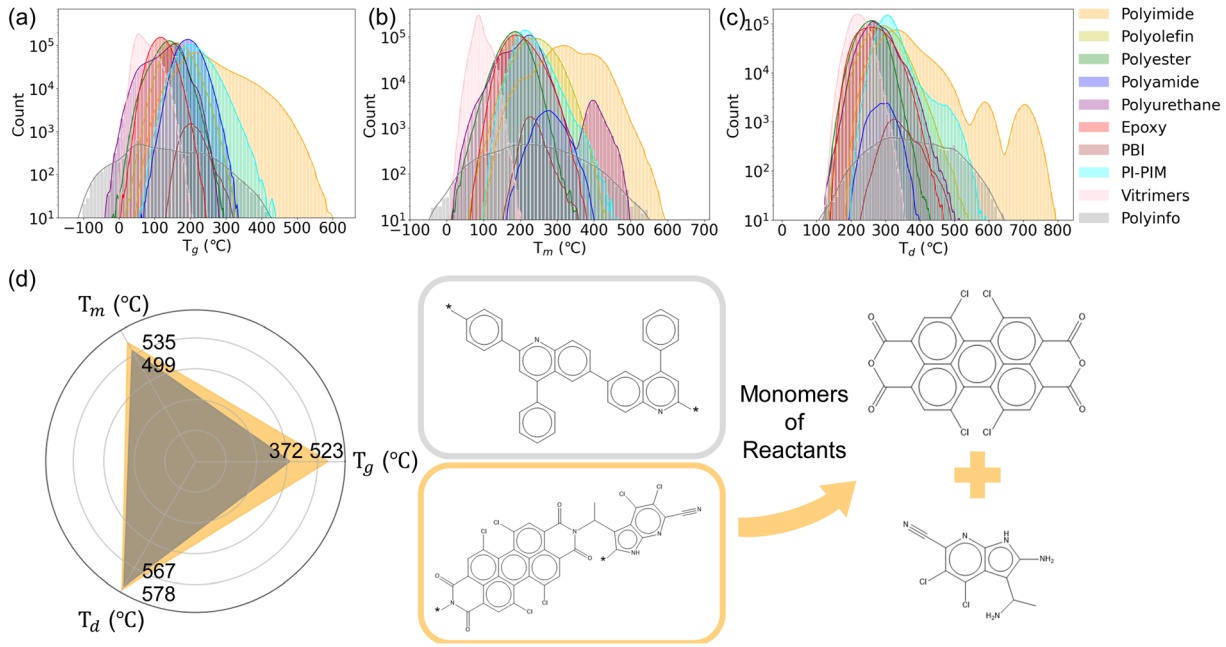

Fig. 5 Distributions of (a) T_g, (b) T_m, and (c) T_d prediction results of real polymers from the PolyInfo dataset and each kind of hypothetical polymer. (d) The real polymer with the highest predicted T_g, T_m, and T_d values in the PolyInfo dataset, and a hypothetical polyimide with predicted T_g, T_m, and T_d values exceeding this performance, along with the small molecule compounds used for its synthesis.

introduction of fused aromatic rings and an increase in the number of chlorine atoms.

2.4.2 Mechanical properties. Mechanical properties are crucial because they determine how a polymer material responds to various forces and stresses, directly influencing its suitability for different applications. Key mechanical properties, such as E, σ_y, and σ_b, provide insights into the material's stiffness, elasticity, and overall durability. E is a measure of a material's elastic properties, defined as the ratio of stress to strain within the elastic deformation range. A higher E value indicates that the material resists deformation more effectively under small strains, exhibiting a stiffer or harder characteristic. σ_y of a polymer is the maximum stress that the material can withstand before yielding occurs. Yielding refers to the point at which the material transitions from elastic deformation to plastic deformation. Beyond the yield strength, the material undergoes irreversible deformation and cannot return to its original shape. Most engineering materials have a strong correlation between σ_y and E.

σ_b of a material is the maximum stress it can endure before failure or fracture occurs. When the applied stress reaches the σ_b, the material will break or fail. σ_y is the level of stress at which a material begins to undergo plastic deformation and is typically lower than σ_b. This is because, after yielding, the material can still withstand additional stress until it ultimately fractures or fails.

These properties are essential for ensuring the material can withstand mechanical loads without deforming or failing, making them vital for applications in the construction, automotive, aerospace, and other industries where structural integrity and performance under stress are critical. Understanding and optimizing mechanical properties enable the development of materials that meet specific performance requirements, enhancing safety, reliability, and functionality in their intended applications.

Fig. 6(a)–(c) display the distribution of E, σ_y, and σ_b prediction values for real polymers from PolyInfo and for each type of generated hypothetical polymer. The overall distribution is similar to that of the thermal properties, with each type of polymer exhibiting a nearly normal distribution. A detailed analysis of each polymer's performance reveals that polyimide continues to demonstrate significant potential, consistent with our previous findings. Additionally, we observed that PI-PIM also shows promising results, particularly in the predicted values for σ_y, and σ_b. PI-PIM is a class of polymers that combine the advantageous properties of polyimides and imine-based polymers. These materials are known for their unique combination of thermal stability, mechanical strength, and chemical resistance, making them highly suitable for various advanced applications. Because of the dynamic nature of the imine bond, the resulting PIM–PIs are malleable, rehealable, and recyclable. The mechanical and thermal properties can be fine-tuned by varying the monomer structures. The study demonstrated that using more rigid monomer precursors, primarily determined by the amine moiety in the imide, resulted in better mechanical performance.[44]

Fig. 6(d) showcases two polymer structures: the real polymer from the PolyInfo dataset with the highest combined predicted values of E, σ_y and σ_b (highlighted within the gray box), and the top-performing hypothetical polymer structure from the 1 million generated hypothetical polyimide structures. The radar chart compares the predicted performance of both polymers,

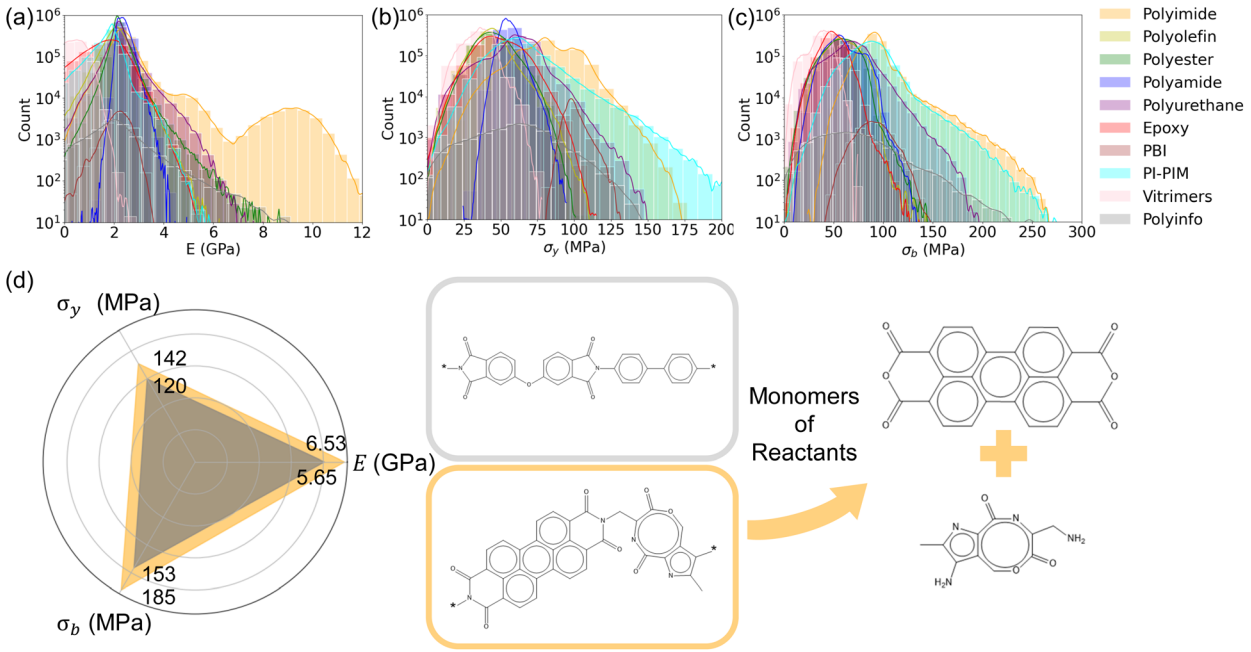

Fig. 6 Distributions of (a) E, (b) σ_y, and (c) σ_b prediction results of real polymers from the PolyInfo dataset and each kind of generated hypothetical polymer. (d) The real polymer with the highest predicted E, σ_y, and σ_b values in the PolyInfo dataset, and a hypothetical polyimide with predicted E, σ_y, and σ_b values exceeding this performance, along with the small molecule compounds used for its synthesis.

while the far right of the figure presents the small molecule compounds used to synthesize the hypothetical polyimide. This comparison clearly demonstrates that the hypothetical polyimide structure outperforms the real polymer in all evaluated aspects, underscoring the significant potential of these newly generated hypothetical polymer structures. Similarly, SHAP was employed to evaluate the impact of substructures on the E and σ_y of the hypothetical polyimide (ESI Fig. S6 and S7†). The SHAP analysis revealed that the high E value of the promising structure is primarily attributed to the introduction of fused aromatic rings and the absence of oxygen atoms bonded to the carbon atoms on the phenyl rings *via* carbon-oxygen single bonds. In addition to these two factors, the high σ_y value of the promising structure is also attributed to the high number of carbon-oxygen double bonds and nitrogen atoms.

Developing structure–function relationships for polymeric materials is inherently challenging due to the need to balance competing properties. For instance, increasing a polymer's strength or stiffness often reduces its flexibility or impact resistance, while enhancing thermal stability may adversely affect processability or toughness. These trade-offs are driven by complex interactions between the molecular structure, morphology, and external conditions, which can simultaneously influence multiple properties. Therefore, designing polymer structures that meet multiple property requirements is a highly challenging task, such as designing a polyimide with high T_g, E and σ_s.[39] This highlights the necessity of having a large hypothetical polymer library to serve as the design space.

2.4.3 Gas permeability. Polymer membranes offer a versatile, cost-effective, and easily processable solution for various separation technologies that play vital roles in addressing climate change (*e.g.*, carbon capture) and enhancing resilience (*e.g.*, water treatment). In gas separation, polymer membranes are extensively utilized in numerous industrial processes such as oxygen enrichment, biogas purification,[66] and post-combustion carbon capture.[67] Carbon capture, in particular, is gaining significant attention as a means to reduce environmental emissions. Membrane technologies are advantageous for their high energy efficiency and operational simplicity, owing to their flexibility and scalability.[68] Key separation processes in different combustion processes—CO_2/N_2 in post-combustion, CO_2/H_2 in pre-combustion, and O_2/N_2 in oxy-combustion—are critical for environmental conservation.[55]

In membrane-based gas separation, a gas mixture is typically driven through a membrane by applying pressure, and separation is achieved due to differences in the permeabilities of the individual gases.[69] The performance of these membrane processes is primarily determined by the membrane's permeability for a specific gas species, denoted as P_i, where i specifies the type of gas. When evaluating the performance of separating gas A from gas B, another crucial measure is the membrane's selectivity, α, defined as $\alpha = P_A/P_B$. An ideal membrane for a particular binary gas separation would exhibit both high permeability and high selectivity. Enhancing gas permeability and selectivity in these membranes would lead to more efficient industrial processes by increasing throughput, reducing energy costs, and achieving a purer product.[70,71] However, there exists a well-known trade-off between permeability and selectivity for polymer gas separation membranes, delineated by the Robeson upper bound.[72]

It is important to note that not all types of polymers are suitable for gas separation. Therefore, in this section, we

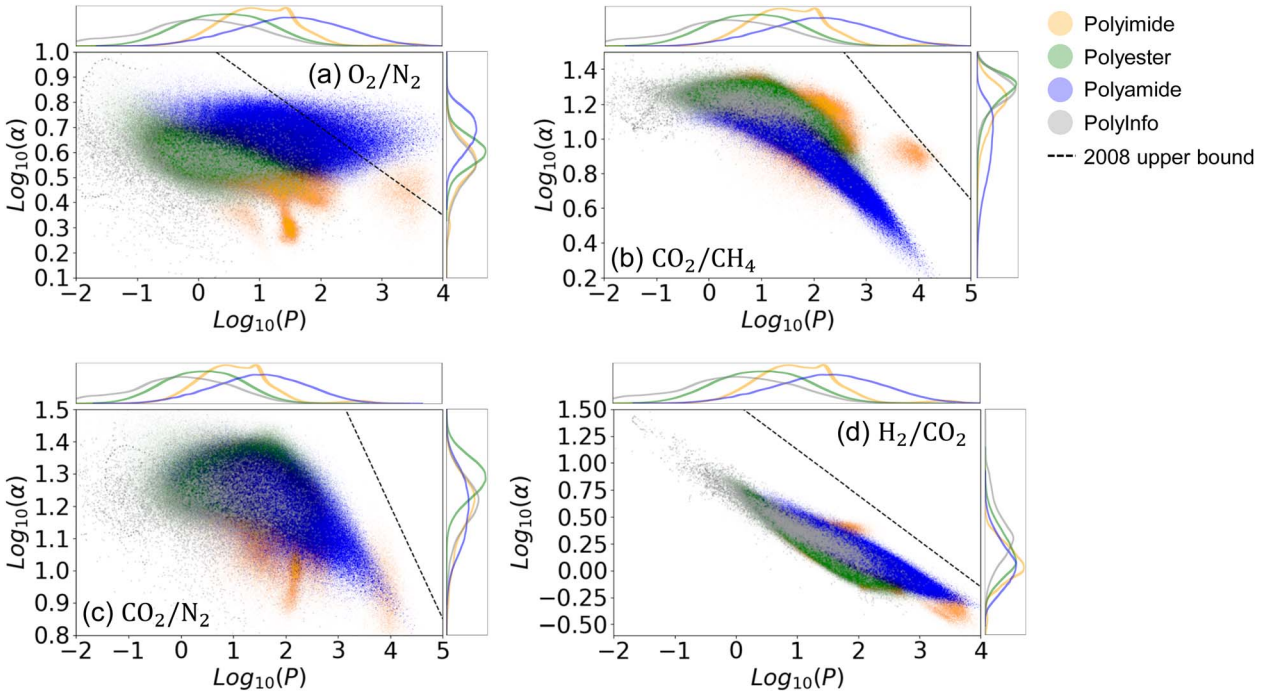

Fig. 7 Visualization of predicted gas permeabilities for real polymers from PolyInfo and hypothetical polymers, including polyimide, polyester, and polyamide. The data are visualized for the following separation processes: (a) O_2/N_2, (b) CO_2/CH_4, (c) CO_2/N_2, and (d) H_2/CO_2. Dashed lines represent the updated 2008 values of the Robeson upper bound. Units of permeability are given in Barrers.

considered only polyimide, polyester, polyamide, and real polymers from the PolyInfo dataset. The predicted permeabilities of these types of hypothetical and real polymers are plotted for O_2/N_2, CO_2/CH_4, CO_2/N_2, and H_2/CO_2 separation in Fig. 7. We can see that in the predictions for all four types of gas separation processes, different types of hypothetical polymers exhibit varying performances across the different gas pairs. The predicted results for hypothetical polyimides, polyesters, and polyamides include many structures that are closer to the Robeson upper bound compared to the real polymers from PolyInfo. As shown in Fig. 7(a), numerous hypothetical polyimides and polyamides even surpass the 2008 values of the Robeson upper bound. Similarly, as shown in Fig. 7(b), some hypothetical polyimides exceed the Robeson upper bound. This demonstrates that our generated hypothetical polymer structures not only have significant potential for developing high-performance materials in terms of thermal and mechanical properties, but they also offer substantial benefits for applications such as gas separation. These polymer structures can greatly assist researchers in advancing separation technologies for natural gas processing, hydrogen production and purification, carbon capture and storage, biogas upgrading, *etc.*

3. Conclusions and outlook

In this study, we successfully demonstrated the generation, analysis, and prediction of properties for a vast array of hypothetical polymer structures, leveraging advances in polymer informatics and ML techniques. Hundreds of quadrillions of hypothetical polymer structures can be generated using small compounds from the GDB-17, GDB-13, and PubChem datasets, combined with well-defined polymerization reactions. We generated millions of hypothetical polymer structures across various classes, including polyimides, polyolefins, polyesters, polyamides, polyurethanes, epoxies, PBIs, vitrimers, and PI-PIMs. The TSNE plot shows that the structures of each polymer type are relatively clustered in the chemical space, with each type generally occupying a specific region.

Through the prediction of glass transition temperature, melting temperature, and decomposition temperature, we identified hypothetical polyimide structures that surpass the highest-performing real polymers, demonstrating significant potential for high-temperature applications. The prediction of Young's modulus, yield strength, and breaking strength revealed that hypothetical polyimides and PI-PIMs exhibit superior mechanical performance compared to existing real polymers, indicating their suitability for demanding applications requiring high strength and durability. The evaluation of gas permeabilities for separation processes such as O_2/N_2, CO_2/CH_4, CO_2/N_2, and H_2/CO_2 showed that many hypothetical polyimides and polyamides approach or exceed the Robeson upper bound, highlighting their potential for efficient gas separation technologies.

The comprehensive analysis and high-throughput screening conducted in this study showcase the immense potential of data-driven methods in polymer science. By identifying high-performance hypothetical polymers, we pave the way for future experimental validation and the development of new materials with tailored properties for specific applications. This

research not only advances our understanding of polymer properties but also provides a valuable open resource database for the scientific community, fostering innovation in materials design and application.

4. Computational methods

4.1 Hypothetical polymer structure generation

The generation of all types of hypothetical polymer structures was implemented using Python and relevant toolkits. For polyimide, polyolefin, polyester, polyamide, and polyurethane, the hypothetical polymer structures were generated using the SMiPoly toolkit. For epoxy, PBI, vitrimers, and PI-PIM, the hypothetical polymer structures were synthesized using the RDKit toolkit.

4.2 Machine learning model

For the FNN model used for T_g, T_m and T_d prediction, the MF with frequency was employed for polymer feature representation. This method identifies substructures within a circle of radius R_M and assigns each substructure a numerical identifier. In this study, the p-SMILES notation of the repeat unit for each sample was utilized, and the fingerprint algorithm was implemented in RDKit with R_M set to 3. A total of 8831 substructures were detected, but to reduce the dimensionality of the input vectors for the FNN model, only 1176 prominent substructures shared by most polymers were retained. An ensemble model, which averages the predictions of twelve FNN models, was used to achieve better prediction performance. The T_g model was optimized through hyperparameter tuning to include four hidden layers with 256, 64, 2048, and 512 neurons, respectively. The T_m model was optimized through hyperparameter tuning to include four hidden layers with 256, 32, 1024, and 1024 neurons, respectively. The T_d model was optimized through hyperparameter tuning to include four hidden layers with 32, 32, 512, and 256 neurons, respectively.

For predicting E, σ_y, and σ_b prediction, the FNN model utilized the MF with frequency for feature representation, with R_M set to 3. Out of a total of 8831 detected substructures, only 129 prominent substructures shared by most polymers were retained to reduce the dimensionality of the input vectors. For each polymer, vectors were created where each bit represents the presence of a detected substructure. An ensemble model, averaging the predictions of twelve FNN models, was employed to enhance prediction performance. Specifically, the model for E was optimized to include a single hidden layer with 40 neurons. The model for σ_y was optimized to have four hidden layers with 8, 8, 8, and 16 neurons, respectively. The model for σ_b was optimized with four hidden layers containing 16, 512, 512, and 1024 neurons, respectively.

For predicting gas permeabilities, the FNN model utilized the MF with frequency for feature representation, with R_M set to 3. From a total of 3209 detected substructures, only 114 prominent substructures shared by most polymers were retained to reduce the dimensionality of the input vectors. The models were optimized with five hidden layers containing 64, 64, 32, 16, and 8 nodes, respectively. The details of the training for all models and the datasets are provided in the ESI† titled "Details of Network Training and Dataset."

Data availability

The categorized small molecule datasets used in this study for hypothetical polymer structure generations can be found at: **https://github.com/ytl0410/PolyUniverse**. The codes for the generation of polyimide, polyolefin, polyester, polyamide, and polyurethane can be found at: **https://github.com/PEJpOhno/SMiPoly**. The codes for the generation of epoxy, PBI, vitrimer, and PI-PIM can be found at: **https://github.com/ytl0410/PolyUniverse**.

Author contributions

Tianle Yue: conceptualization, data curation, formal analysis, investigation, methodology, resources, visualization, writing – original draft, writing – review & editing. Jianxin He: formal analysis, investigation, methodology, writing – original draft. Ying Li: conceptualization, formal analysis, funding acquisition, project administration, investigation, methodology, supervision, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We gratefully acknowledge financial support from the Air Force Office of Scientific Research through the Air Force's Young Investigator Research Program (FA9550-20-1-0183; Program Manager: Dr Ming-Jen Pan and Capt. Derek Barbee), Air Force Research Laboratory/UES Inc. (FA8650-20-S-5008, PICASSO program), and the National Science Foundation (CMMI-2332276, CMMI-2316200, and CAREER-2323108). Y. L. would also like to acknowledge the support from 3M's Non-Tenured Faculty Award. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Defense or the National Science Foundation. Support for this research was also provided by the University of Wisconsin-Madison, Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

References

1 C. S. Brazel and S. L. Rosen, *Fundamental Principles of Polymeric Materials*, John Wiley & Sons, 2012.
2 T. Lei, J.-Y. Wang and J. Pei, Roles of flexible chains in organic semiconducting materials, *Chem. Mater.*, 2014, **26**(1), 594–603.
3 M. A. F. Afzal, *From Virtual High-Throughput Screening and Machine Learning to the Discovery and Rational Design of*

Polymers for Optical Applications, State University of New York at Buffalo, 2018.
4 A. S. Abd-El-Aziz, M. Antonietti, C. Barner-Kowollik, W. H. Binder, A. Böker, C. Boyer, M. R. Buchmeiser, S. Z. Cheng, F. D'Agosto and G. Floudas, The next 100 years of polymer science, *Macromol. Chem. Phys.*, 2020, **221**(16), 2000216.
5 J. R. Fried, *Polymer Science and Technology*, Pearson Education, 2014.
6 R. G. Parr, Density functional theory, in *Electron Distributions and the Chemical Bond*, Springer, 1982, pp. 95–100.
7 A. J. Cohen, P. Mori-Sánchez and W. Yang, Insights into current limitations of density functional theory, *Science*, 2008, **321**(5890), 792–794.
8 D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Elsevier, 2023.
9 D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, Cambridge University Press, 2004.
10 A. Agrawal and A. Choudhary, Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, *APL Mater.*, 2016, **4**(5), 053208.
11 Y. Gil, M. Greaves, J. Hendler and H. Hirsh, Amplify scientific discovery with artificial intelligence, *Science*, 2014, **346**(6206), 171–172.
12 B. Sanchez-Lengeling and A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering, *Science*, 2018, **361**(6400), 360–365.
13 G. Chen, Z. Shen, A. Iyer, U. F. Ghumman, S. Tang, J. Bi, W. Chen and Y. Li, Machine-learning-assisted de novo design of organic molecules and polymers: opportunities and challenges, *Polymers*, 2020, **12**(1), 163.
14 K. Sattari, Y. Xie and J. Lin, Data-driven algorithms for inverse design of polymers, *Soft Matter*, 2021, **17**(33), 7607–7622.
15 R. Batra, L. Song and R. Ramprasad, Emerging materials intelligence ecosystems propelled by machine learning, *Nat. Rev. Mater.*, 2021, **6**(8), 655–678.
16 L. Chen, G. Pilania, R. Batra, T. D. Huan, C. Kim, C. Kuenneth and R. Ramprasad, Polymer informatics: Current status and critical next steps, *Mater. Sci. Eng., R*, 2021, **144**, 100595.
17 C. Kim, A. Chandrasekaran, T. D. Huan, D. Das and R. Ramprasad, Polymer genome: a data-powered polymer informatics platform for property predictions, *J. Phys. Chem. C*, 2018, **122**(31), 17575–17585.
18 S. Wu, Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu and J. Shiomi, Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm, *npj Comput. Mater.*, 2019, **5**(1), 66.
19 L. A. Miccio and G. A. Schwartz, From chemical structure to quantitative polymer properties prediction through convolutional neural networks, *Polymer*, 2020, **193**, 122341.
20 L. A. Miccio and G. A. Schwartz, Localizing and quantifying the intra-monomer contributions to the glass transition temperature using artificial neural networks, *Polymer*, 2020, **203**, 122786.
21 L. Ning, Artificial neural network prediction of glass transition temperature of fluorine-containing polybenzoxazoles, *J. Mater. Sci.*, 2009, **44**, 3156–3164.
22 W. Liu, Prediction of glass transition temperatures of aromatic heterocyclic polyimides using an ANN model, *Polym. Eng. Sci.*, 2010, **50**(8), 1547–1557.
23 D. Palomba, G. E. Vazquez and M. F. Díaz, Novel descriptors from main and side chains of high-molecular-weight polymers applied to prediction of glass transition temperatures, *J. Mol. Graphics Modell.*, 2012, **38**, 137–147.
24 B. E. Mattioni and P. C. Jurs, Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks, *J. Chem. Inf. Comput. Sci.*, 2002, **42**(2), 232–240.
25 W. Liu and C. Cao, Artificial neural network prediction of glass transition temperature of polymers, *Colloid Polym. Sci.*, 2009, **287**, 811–818.
26 J. F. Pei, C. Z. Cai, Y. M. Zhu and B. Yan, Modeling and Predicting the Glass Transition Temperature of Polymethacrylates Based on Quantum Chemical Descriptors by Using Hybrid PSO-SVR, *Macromol. Theory Simul.*, 2013, **22**(1), 52–60.
27 C. Higuchi, D. Horvath, G. Marcou, K. Yoshizawa and A. Varnek, Prediction of the glass-transition temperatures of linear homo/heteropolymers and cross-linked epoxy resins, *ACS Appl. Polym. Mater.*, 2019, **1**(6), 1430–1442.
28 G. Pilania, C. N. Iverson, T. Lookman and B. L. Marrone, Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers, *J. Chem. Inf. Model.*, 2019, **59**(12), 5013–5025.
29 S. Goswami, R. Ghosh, A. Neog and B. Das, Deep learning based approach for prediction of glass transition temperature in polymers, *Mater. Today: Proc.*, 2021, **46**, 5838–5843.
30 L. A. Miccio and G. A. Schwartz, Mapping chemical structure–glass transition temperature relationship through artificial intelligence, *Macromolecules*, 2021, **54**(4), 1811–1817.
31 A. Patra, R. Batra, A. Chandrasekaran, C. Kim, T. D. Huan and R. Ramprasad, A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap, *Comput. Mater. Sci.*, 2020, **172**, 109286.
32 L. Chen, C. Kim, R. Batra, J. P. Lightstone, C. Wu, Z. Li, A. A. Deshmukh, Y. Wang, H. D. Tran and P. Vashishta, Frequency-dependent dielectric constant prediction of polymers using machine learning, *npj Comput. Mater.*, 2020, **6**(1), 61.
33 J. P. Lightstone, L. Chen, C. Kim, R. Batra and R. Ramprasad, Refractive index prediction models for polymers using machine learning, *J. Appl. Phys.*, 2020, **127**(21), 215105.
34 S. Otsuka, I. Kuwajima, J. Hosoya, Y. Xu and M. Yamazaki, PoLyInfo: Polymer database for polymeric materials design, in *2011 International Conference on Emerging Intelligent Data and Web Technologies*, IEEE, 2011, pp. 22–29.

35 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen and B. Yu, PubChem 2019 update: improved access to chemical data, *Nucleic Acids Res.*, 2019, **47**(D1), D1102–D1109.
36 L. C. Blum and J.-L. Reymond, 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13, *J. Am. Chem. Soc.*, 2009, **131**(25), 8732–8733.
37 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, *J. Chem. Inf. Model.*, 2012, **52**(11), 2864–2875.
38 R. Ma and T. Luo, PI1M: a benchmark database for polymer informatics, *J. Chem. Inf. Model.*, 2020, **60**(10), 4684–4690.
39 L. Tao, J. He, N. E. Munyaneza, V. Varshney, W. Chen, G. Liu and Y. Li, Discovery of multi-functional polyimides through high-throughput screening using explainable machine learning, *Chem. Eng. J.*, 2023, **465**, 142949.
40 R. Wang, Y. Zhu, J. Fu, M. Yang, Z. Ran, J. Li, M. Li, J. Hu, J. He and Q. Li, Designing tailored combinations of structural units in polymer dielectrics for high-temperature capacitive energy storage, *Nat. Commun.*, 2023, **14**(1), 2406.
41 S. Kim, C. M. Schroeder and N. E. Jackson, Open macromolecular genome: Generative design of synthetically accessible polymers, *ACS Polym. Au*, 2023, **3**(4), 318–330.
42 M. Ohno, Y. Hayashi, Q. Zhang, Y. Kaneko and R. Yoshida, SMiPoly: Generation of a Synthesizable Polymer Virtual Library Using Rule-Based Polymerization Reactions, *J. Chem. Inf. Model.*, 2023, **63**(17), 5539–5548.
43 B. S. Ferrari, M. Manica, R. Giro, T. Laino and M. B. Steiner, Predicting polymerization reactions via transfer learning using chemical language models, *npj Comput. Mater.*, 2024, **10**(1), 119.
44 X. Shen, Y. Ma, S. Luo, R. Tao, D. An, X. Wei, Y. Jin, L. Qiu and W. Zhang, Malleable and recyclable imide–imine hybrid thermosets: influence of imide structure on material property, *Mater. Adv.*, 2021, **2**(13), 4333–4338.
45 C. H. Chan, J.-T. Chen, W. S. Farrell, C. M. Fellows, D. J. Keddie, C. K. Luscombe, J. B. Matson, J. Merna, G. Moad and G. T. Russell, Reconsidering terms for mechanisms of polymer growth: the "step-growth" and "chain-growth" dilemma, *Polym. Chem.*, 2022, **13**(16), 2262–2270.
46 G. Odian, *Principles of Polymerization*, John Wiley & Sons, 2004.
47 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich and B. Al-Lazikani, ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res.*, 2012, **40**(D1), D1100–D1107.
48 T. Sterling and J. J. Irwin, ZINC 15–ligand discovery for everyone, *J. Chem. Inf. Model.*, 2015, **55**(11), 2324–2337.
49 H. E. Pence and A. Williams, *ChemSpider: An Online Chemical Information Resource*, ACS Publications, 2010.
50 C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin and S. A. Strawbridge, DrugBank 6.0: the DrugBank knowledgebase for 2024, *Nucleic Acids Res.*, 2024, **52**(D1), D1265–D1275.
51 L. Van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.*, 2008, **9**(11), 2579–2605.
52 L. Tao, V. Varshney and Y. Li, Benchmarking machine learning models for polymer informatics: an example of glass transition temperature, *J. Chem. Inf. Model.*, 2021, **61**(11), 5395–5413.
53 L. Tao, G. Chen and Y. Li, Machine learning discovery of high-temperature polymers, *Patterns*, 2021, **2**(4), 100225.
54 T. Yue, J. He, L. Tao and Y. Li, High-throughput screening and prediction of high modulus of resilience polymers using explainable machine learning, *J. Chem. Theory Comput.*, 2023, **19**(14), 4641–4653.
55 J. Yang, L. Tao, J. He, J. R. McCutcheon and Y. Li, Machine learning enables interpretable discovery of innovative polymers for gas separation membranes, *Sci. Adv.*, 2022, **8**(29), eabn9545.
56 G. Landrum, RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling, *Greg Landrum*, 2013, **8**(31), 5281.
57 D. W. Van Krevelen and K. Te Nijenhuis, *Properties of Polymers: Their Correlation with Chemical Structure; Their Numerical Estimation and Prediction from Additive Group Contributions*, Elsevier, 2009.
58 M. M. Coleman, *Fundamentals of Polymer Science: An Introductory Text*, Routledge, 2019.
59 S. Diaham, Polyimide in electronics: applications and processability overview, *Polyimide Electron. Electr. Eng. Appl.*, 2021, 2020–2021.
60 Y. S. Negi, S. R. Damkale and S. Ansari, Photosensitive polyimides, *J. Macromol. Sci., Part C: Polym. Rev.*, 2001, **41**(1–2), 119–138.
61 M. Hasegawa, N. Sensui, Y. Shindo and R. Yokota, Structure and Properties of Novel Asymmetric Biphenyl Type Polyimides, *J. Photopolym. Sci. Technol.*, 1996, **9**(2), 367–378.
62 I. Gouzman, E. Grossman, R. Verker, N. Atar, A. Bolker and N. Eliaz, Advances in polyimide-based materials for space applications, *Adv. Mater.*, 2019, **31**(18), 1807738.
63 S. Ghaffari-Mosanenzadeh, O. A. Tafreshi, S. Karamikamkar, Z. Saadatnia, E. Rad, M. Meysami and H. E. Naguib, Recent advances in tailoring and improving the properties of polyimide aerogels and their application, *Adv. Colloid Interface Sci.*, 2022, **304**, 102646.
64 E. P. Favvas, F. K. Katsaros, S. K. Papageorgiou, A. A. Sapalidis and A. C. Mitropoulos, A review of the latest development of polyimide based membranes for CO_2 separations, *React. Funct. Polym.*, 2017, **120**, 104–130.
65 S. Lundberg, *A unified approach to interpreting model predictions*, *arXiv*, 2017, preprint, arXiv:1705.07874, DOI: **10.48550/arXiv.1705.07874**.
66 S. Basu, A. L. Khan, A. Cano-Odena, C. Liu and I. F. Vankelecom, Membrane-based technologies for biogas separations, *Chem. Soc. Rev.*, 2010, **39**(2), 750–768.
67 S. Zhao, P. H. Feron, L. Deng, E. Favre, E. Chabanon, S. Yan, J. Hou, V. Chen and H. Qi, Status and progress of membrane contactors in post-combustion carbon capture: A state-of-

the-art review of new developments, *J. Membr. Sci.*, 2016, **511**, 180–206.

68 Y. Han and W. W. Ho, Polymeric membranes for CO_2 separation and capture, *J. Membr. Sci.*, 2021, **628**, 119244.

69 B. D. Freeman, Basis of permeability/selectivity tradeoff relations in polymeric gas separation membranes, *Macromolecules*, 1999, **32**(2), 375–380.

70 D. F. Sanders, Z. P. Smith, R. Guo, L. M. Robeson, J. E. McGrath, D. R. Paul and B. D. Freeman, Energy-efficient polymeric gas separation membranes for a sustainable future: A review, *Polymer*, 2013, **54**(18), 4729–4761.

71 H. B. Park, J. Kamcev, L. M. Robeson, M. Elimelech and B. D. Freeman, Maximizing the right stuff: The trade-off between membrane permeability and selectivity, *Science*, 2017, **356**(6343), eaab0530.

72 L. M. Robeson, The upper bound revisited, *J. Membr. Sci.*, 2008, **320**(1–2), 390–400.